SERVICE CENTER IMPLEMENTATION TEAM (SCIT)

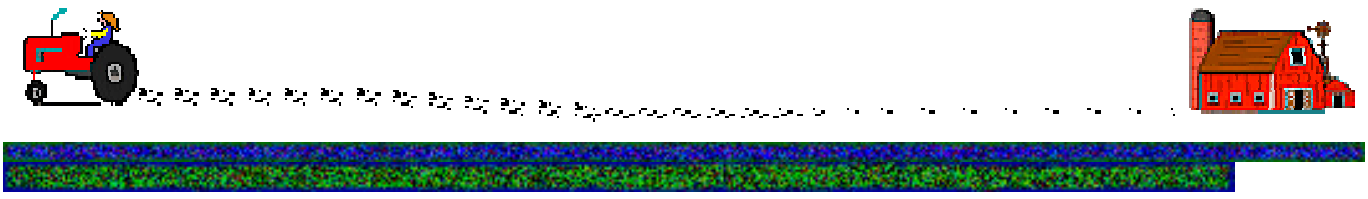
DATA WAREHOUSE STRATEGY DOCUMENT

Submitted to:

U.S. Department of Agriculture,
OCIO/CCE

Submitted by:

USDA
Farm Service Agency
Natural Resources Conservation Service
Rural Development



Executive Summary

The United States Department of Agriculture agencies, specifically, business personnel from the Natural Resources Conservation Service, Farm Service Agency, and Rural Development would benefit from the ability to more easily obtain comprehensive, accurate, and timely information relevant to their respective business operations. The capability to extract information and analyze it has long been a problem area for these agencies, as the information is maintained across a wide range of legacy systems that were not designed to present consolidated views or to support information accessibility or decision-making. **For the purposes of this strategy document, USDA is synonymous with only the three partner agencies.**

Perhaps the most important requirement for a data warehouse (DW) is that it delivers a solution to a well-defined business problem. Without a clear focus defined by the business, a warehouse risks becoming a technology initiative only and its chances for survival get greatly reduced. It's also important that the project have a specific goal with tangible benefits (financial and otherwise) that can be measured. Measuring the success of a data warehouse helps ensure the project will get continued funding.

The ability to "slice and dice" massive amounts of data maintained in non-integrated legacy systems is more important today than ever. Availability of intelligent information in an expedient manner is absolutely essential for the USDA decision-makers to sustain their respective program missions. Operating within austere budget levels, reduced personnel strengths, and an ever-spiraling workload, USDA's business managers must continue to properly administer a wide range of loan, farm, and conservation programs (e.g. Single Family Housing, Water and Waste, Telemedicine and Long-distance Learning, Multi-Family Housing, Community Facilities, Business and Industry, Integrated Accountability System, Workload Assessment, Workload Measurement, Resource Conservation and Development, Watershed, Core Accounting System etc.). The availability of a data warehouse is essential to effectively and efficiently provide this critical business data.

The primary goal of this initiative is to provide strategic information so that USDA's business managers can continue to thrive in today's environment where federal agencies must do more with less and still improve customer services. The decision makers must discover which programs are working and which are not so that remedial actions can be quickly implemented in a cost effective and efficient manner that prevents significant losses to the U.S. taxpayer and disruption of service to rural America. This initiative will provide a unified view of disparate databases, assisting partner agency managers to get a better handle on the enormous amounts of data that support their daily activities.

A comprehensive data warehouse will enable managers to look for patterns of defaults on loans, monitor performance of lenders from the private sector engaged in making and servicing guaranteed loans, and by collecting metrics over time, predict when problems might occur and facilitate the planning of preventive measures. Other objectives of the USDA data warehouse are:

1. Ensure that information about natural resource conditions and management practices are in an appropriate, usable form and accessible to local users at scales that range from fields to farms to watersheds or ecosystems.
2. Ensure that policy makers and citizens have reliable, timely, regional, and national information on natural resource status and trends.
3. Provide a resource for USDA partner agencies to increase productivity to accomplish the mission and reduce maintenance costs.
4. Enable the partner agencies to better fulfil its commitments of the Government Performance and Results Act, which strives to make government maximize operational efficiencies and minimize costs.
5. Improve the quality, reliability, and consistency of data that facilitates better and more accurate business decisions.

It is the objective of the USDA partner agencies to create their DW as a place where decision-makers and other users can explore data limitlessly, run ad-hoc queries and drill for details at will. The data will be subject/oriented, integrated, time-variant and non-volatile. As such, building the USDA data warehouse will involve the following 4-step process:

1. Extract, model, and assemble data from the operational legacy systems;
2. Transform operational data to a user focused view;
3. Distribute and manage data changes to the warehouse, and
4. Provide access to the data through Decision Support Systems (DSS) or Executive Information System (EIS) tools.

This document outlines the requirements for establishment of a USDA partner agency data warehousing environment. Acquisition is designed to cover the full range of extraction, cleansing, transformation, and accessibility tools. Proposed acquisition will be in line with the CCE (Common Computing Environment) technical architecture. The current IT investment strategy outlines the procurement for enterprise servers and relational database software. Should the necessary CCE funding be unavailable in the near timeframe, individual data mart waivers, which are business area specific and consistent with the data warehouse enterprise strategy, will be requested to keep the current momentum on a forward pace.

From a technical standpoint, easy accessibility to information contained in the data warehouse enables the user to query and format reports that satisfy data calls from a wide variety of sources. This accessibility negates the need to program numerous report modifications, thus freeing up time for the partner agencies' application developers to build new and enhanced business systems that improve operational efficiency.

Although many areas within the USDA can benefit from data warehousing, the most promising will be reduced delinquencies, customer profiling, significant decrease in writeoffs, demand forecasting, improved asset management and cost control, better distribution of funds throughout Rural America, and increased efficiency in sharing data between agencies, such as crop loss disaster assistance program information between FSA and Risk Management Agency (RMA) via WEB access. Other benefits would be improved audit analysis by FSA County Office Reviewers (COR), Office of Management and Budget (OMB), and Office of Inspector General (OIG). USDA would be able to directly provide Environmental Protection Agency (EPA) with record data for environmental modeling projects. These are areas in which the USDA hopes to save program funds that could be utilized to stimulate business and economic conditions, provide clean water to all of Rural America, and give more farm and housing credit where needed in depressed areas.

Table of Contents

Table of Figures

Section 1 Introduction

1.1 Background

The United States Department of Agriculture agencies, including the Natural Resources Conservation Service (NRCS), Farm Service Agency (FSA), and Rural Development (RD), recognize the need for timely access to comprehensive, accurate, timely and relevant information on which to base important business decisions. It is more important today than ever for business personnel to have the capability to "slice and dice" massive quantities of data that is currently available only through disparate legacy systems. The establishment of a data warehouse-oriented architecture would significantly improve the information availability in support of a wide range of loan, farm and conservation programs (such as Single Family Housing, Community Facilities, Business and Industry, Tobacco, Accounting, Conservation Reserves, Price Support, Commodity, Loan Deficiencies, etc.)

The USDA Service Center Data Management Team (SCDT) is in the process of establishing the processes and infrastructure through which data (tabular and spatial) is developed and managed so as to protect the value of the data asset, ensure data quality, improve data organization and enhance the ability to share and reuse data. The Data warehouses environment will contain key elements in this infrastructure. This environment will provide the ability to integrate critical data across partner agencies and geographic boundaries to support research, trend analysis and executive decision-making. Data marts within this data warehouse environment will provide a single, consolidated source for customer information across the three partner agencies, resulting in a substantially greater business value for the USDA partner agencies.

A charter for the Service Center Data Management Team (SCDT) Data Warehouse Technical Working Group (TWG) was developed in July 1998. It outlines the team's responsibility, which is to

> *"Provide the leadership for the strategy, framework, and common technical solutions essential for the successful development and deployment of partner agencies' data warehouses*[1]*. "*

The goal of the SCDT Data Warehouse TWG is to develop an enterprise strategy for incremental fielding of business-driven data marts. This strategy includes the establishment of a common data warehousing architecture, as well as adopting a standard set of tools to support data warehouse development and maintenance and to support business-user accessibility.

[1] The Service Center Data Management Team (SCDT) Data Warehouse Technical Working Group (TWG) Charter, August 19, 1998.

1.2 Authority & Responsibility

The Data Warehouse TWG receives direction and guidance from the Service Center Data Management Steering Body, which is comprised of partner agencies' Chief Information Officers (CIO) and National Food and Agriculture Council (NFAC) Executive Officers. The Steering Body approves the charter, team members, and all final deliverables. The Steering Body is briefed periodically on team activities and plans. The Data Warehouse TWG is a technical working group of the Service Center Data Team. The Data Warehouse TWG project plan is a subsection of the overall Data Management project plan.

The Data Warehouse TWG has the authority to manage the implementation of the data warehouse initiative. It works in close coordination with the Business Executive Sponsors, Subject Matter Experts and selected Business Analysts from the three Service Center partner agencies. It operates within established Departmental guidelines and service center agencies' budgetary limits.

The responsibilities of the Data Warehouse TWG include:

- facilitating the identification of business needs,
- obtaining sponsorship at the appropriate levels,
- recognizing key cultural issues,
- managing expectations,
- budgeting proposed activities for authorization and funding,
- building the common architecture that takes into account the need for scaleable elements,
- determining initial subject areas,
- selecting development and operational support tools, and
- developing an implementation strategy for an Enterprise-wide DW initiative.

The responsibilities of this team will be transitioned to the new Information Technology (IT) organization, when this organization becomes operational. The new IT organization will assume responsibility for future implementations, ongoing management and maintenance.

1.3 Assertions & Guiding Principles

The following assertions provide the framework for the successful implementation of data warehousing capability for the partner agencies.

1. Requirements for the Enterprise Data Warehouse initiative will be business driven.
2. The data warehouse initiative will, upon being developed and deployed, provide greater business value (benefit exceeds cost) than if they were not implemented.
3. Each of the data warehouse project(s) require the support of its sponsors. The executive-level staff who approve the project must agree with the first assertion and communicate their belief in this to everyone involved in the project.
4. The data warehouse project(s) require consensus and support among its stakeholders (the staff whose jobs will be most directly effected by the deployment of a warehouse solution).
5. The data warehouse(s) will need to provide and support information needs of multiple, diversified users.

Furthermore, the following principles shall be used to guide the data warehouse project(s) with regard to the existing systems:

1. Existing legacy applications will not be affected by the implementation of the data warehousing project(s).
2. The daily operations of an existing on-line transaction processing systems (OLTP) will not be disrupted by the implementation of the data warehousing project(s).
3. The data required to load into the data warehouse(s) may be extracted from existing OLTP applications.

1.4 Purpose and Scope

This document describes the overall strategy and implementation plan for the development of these business-driven data marts, and their integration based on an enterprise perspective.

It provides a high-level description of:

- data warehouses and data marts;
- the overall concept for integrating data warehouse technology into the IT infrastructure of the USDA;

- the benefits associated with the data warehouse approach;
- infrastructure requirements and selection criteria;
- design strategies;
- the implementation approach.

Its purpose is to provide a high-level understanding of what data warehouses are and what benefits they offer. It also lays out a plan for ensuring that these benefits may be realized within the overall data management goals for the USDA and its partner agencies.

This initiative outlines the high level data warehousing strategy within the context of the USDA data management principles and policies. Implementation of the USDA enterprise data warehouse will be evolutionary and realized through the establishment of incremental architected data marts that conform to an enterprise set of standards, controls, and procedures.

Section 2: Business Benefits of Data Warehousing

The advent of data warehousing technology offers numerous benefits to today's business enterprises, and to the USDA partner agencies in particular. The ability to summarize and integrate vast quantities of historical data across many inter-related subject areas and disparate legacy systems offers powerful research and trend analysis capabilities that can benefit USDA management and executive decision makers by making the "right information available to the right people at the right time".

2.1 General Benefits

Like many large organizations, Farm Service Agency (FSA), Rural Development (RD), and Natural Resource Conservation Service (NRCS) have accumulated information in vast amounts. This information is usually not accessible to business users. Information Technology (IT) professionals must develop solutions to make such data available to the business user. Instead of submitting an extensive list of requested reports to IT, a business user could simply make a specific request at their PC to the data warehouse and see immediate results. IT professionals can use their expertise to select, streamline and organize raw data for the data warehouse instead of spending time and resources creating ad hoc reports. The data warehouse will eliminate a very costly and time consuming report generation process for the partner agencies.

It is envisioned that the data warehouse concept of user generated ad hoc reports will be incorporated into the Service Center Implementation Team's Business Process Reengineering (BPR) initiatives. This will result in a significant decrease in system development time. This equates to lower costs and quicker implementation of BPR processes essential for the future well being of the USDA.

Civil Rights is an issue that cuts across all programs and is an area of potential significant benefits. Agencies would be able to be proactive instead of reactive in the equitable delivery of USDA programs. With data warehousing technology program managers would be able to drill down to the customer level and match that with local demographic information to evaluate trends or out of norm data for closer review. For example, if a particular zip code had an 80% non-minority population and yet 80% of the loan foreclosures were from minority borrowers, an issue of discrimination may exist. Management could then determine the truth and take action. The intangible financial savings would be that the Department would not be in litigation or be subject to payments as a result of discrimination. Data warehousing will facilitate fairness and enhance borrower success by providing timely and accurate information to decision-makers. For example, if in a particular area it is determined that historically under-served Americans lose their homes at a greater rate than other borrowers, the remedial solution could be a direly needed education program. With the ability to marry

demographic and loan information, the cause would be more easily identified and corresponding program decisions made that improve the rate of successful home ownership, specifically in targeted areas.

Loan losses as the result of foreclosures are another area with significant long- and short-term benefits. Currently most loan programs have few Management Information Systems (MIS) tools at their disposal that would help them learn the underlying causes for foreclosures. In many cases, the Department simply pays the losses and then tries to figure out why they occurred. Again, data warehousing would allow program managers to be proactive in the management of losses. For example, many loan factors could be reviewed at the loan level, benchmarks established, and trends monitored for not only the borrower but the lender and servicer as well. If USDA could determine, through tracking, that 90% of all borrowers who had their loan(s) re-amortized went into foreclosure, the agency may decide to limit re-amortization of loans or only allow them under conditions that were identified as being trends from the 10% who were successful. This would result in quicker property disposal and significant corresponding cost savings. Additionally, the personnel engaged in the time consuming process of loan re-amortizing, could be concentrated in other servicing areas to further reduce delinquencies and loan losses.

The secret to lower foreclosures and reduced loan losses is to make good loans. That requires the ability to identify the characteristics of a good loan/borrower so lenders can be told what standards are acceptable to a particular loan program. Since its inception in 1991 the Rural Housing Single Family Guaranteed Program has paid out over $50 million dollars in loan losses on a portfolio worth more than $10 billion dollars. In order for program managers to be able to reduce losses (which currently fall within industry standards), timely and accurate information provided by data warehousing technology is required. Decision-makers would be able to compare loan characteristics of successful borrowers to unsuccessful borrowers; lenders with high losses to lenders with low losses and make policy decisions as trends are identified. Answers to the following type questions could be easily ascertained:

- Are loan to debt and income ratios too low/ high?
- Do failed borrowers have worse credit history?
- Is payment shock a factor (increasing your housing expense by 200% +/-)?

All information needed to make informed decisions would be available quickly and easily to allow program managers the ability to make decisions needed to reduce loan losses. In a portfolio the size of the housing program, these reductions in loan losses would provide significant cost savings.

In the tobacco program the Office of Inspector General (OIG) has identified potential fraud, as it relates to marketing quotas. FSA relies on a paper distribution process that delivers information and reports on sale transactions after the transaction is more than a week old. The Agency is then in the position of playing catch up on potential fraud cases. Once identified these violations may or may not be pursued through litigation. Data warehousing would give the agency the ability to monitor the near real time sale transactions, thus catching potential fraud within a 48 hour period, which could save millions of dollars in lost marketing quota penalty collections.

Data warehousing facilitates the summary and integration of large quantities of legacy data, across data subject areas. This supports research and trend analysis opportunities across disparate legacy systems. This type of analysis ranges from impossible to extremely costly without a data warehouse.

The type, quantity, and quality of data contained in the legacy environment are usually inadequate for decision support. Operational systems contain only the data to meet day-to-day business requirements, while data warehouses contain a large quantity of historical data for informational purposes. The warehouse will be used to analyze trends over a long period of time. There is also a need to merge operational data, external data, and personal data in ad hoc queries. Most operational systems are not positioned to meet this objective.

The data warehouse puts the data more directly into the hands of users. Users are able to perform ad hoc queries and generate reports directly through the use of OLAP tools rather than requiring IT resources to perform these functions.

The data warehouse will provide a single source of data for commonly-used subject areas. This will improve the consistency of that data and reduce overhead to manage replication and synchronization issues across multiple sources. Since data warehousing presents views of data at a documented point in time, confusion over time-related "inconsistencies" in data extracted through OLTPs is eliminated.

The need for standard names and definitions for data in the warehouse facilitates the use of a common language, thereby improving communication across organizations and functional business areas.

The data warehouse reduces the maintenance and operational costs associated with many legacy systems, which may eventually be retired as data is migrated to the warehouse.

2.2 Specific Benefits

2.2.1 Congressional

Accurate, detailed information pertaining to budgeting, funding, demographics and customer data, needs to be readily available for retrieval for Congressional inquiries. The volume of Congressional inquiries continues to increase.

Currently, in order to collect data, the National Office has to call the State Offices, who in turn may have to call field offices, have them collect the data and send it back to the National Office through the State Office. Often the report data is inaccurate and has to be manually verified, or is not in the proper format and has to be manually retyped. This is an inefficient use of time and resources.

2.2.2 Program efficiency & effectiveness

Data Warehousing provides needed information on how to improve program delivery. For instance, in loan programs with the ability to pull loan information and compare it with demographic information, trends could be analyzed and proper conclusions drawn as to what improvements could be made to loan programs in order to insure equitable loan program delivery.

2.2.3 Capital Planning

The availability of summarized and consolidated information through a warehouse would significantly reduce the 2-3 months currently required to compile budget data. This will allow additional time for analysis of that data. It is estimated that the warehouse could reduce the time for data collection by as much as 50%. Likewise, it currently takes 6 months to collect the necessary information and generate a financial statement; the support of a data warehouse is expected to reduce this to 2 weeks.

For example, operations management has expressed concerns that the information provided to OMB is not considered timely. This has resulted in millions of dollars in reductions to the budget due to poor assumptions. The availability of this information through a data warehouse would have improved the turn around of information, and could have precluded the budget reductions.

2.2.4 Performance Measures

Timely, Accurate, and Consolidated Summary Data will be available for program performance measurement of:

- National and State Strategic Plans;
- Government Performance & Results Act of 1993 reporting;

- Personel requirements;
- Information Technology Initiatives;
- Year 2000 compliance.

All of this will be used in measuring the effectiveness of program outcomes and will give decision-makers the tools to make any necessary course corrections.

2.2.5 Oversight Agency Review

Availability of accurate, consistent and timely data will enable senior managers to readily identify program issues. This information will be utilized to make improvements/modifications to programs, assuring compliance with:

- Internal Controls;
- Federal Managers Financial Integrity Act (FMFIA), oversight by OMB, GAO, OIG, and OCFO;
- Debt Collection Improvement Act;
- Guaranteed Lending Review;
- Compliance with Federal statutes, Executive Orders, and Agency Regulations and policies and;
- Performance Appraisals.

2.2.6 Business Processing Reengineering

In order to meet Business Processing Reengineering (BPR) mandates, agencies must be able to evaluate program delivery and access effectiveness, down to the lowest level of information. Reengineering of the following major systems will be accomplished by the partner agencies:

- Core Accounting System (CORE). FSA has determined that it currently costs approximately $15,000.00 to generate a mainframe report; there are currently 250 critical reports required to support CORE. The use of data warehousing technology offers the opportunity to significantly reduce this cost by putting report generation capabilities into the hands of the users rather than requiring intensive support from IT staff;.
- FSA's Farm Loan Program;

- FSA Processed Commodities Inventory Management System;
- FSA Grain Inventory Management System;
- FSA Cotton Inventory Management System. This system is currently in the process of business process reengineering and redevelopment;
- FSA Tobacco Marketing Price Support Program System. This system is currently in the process of business process reengineering and redevelopment;
- Multi-Family Housing Integrated System;
- Guaranteed Loans Fund Reservation System.

Data warehousing complements these reengineering efforts and will allow agencies to obtain management information in multiple consolidated areas. Benefits will include:

- Reduced manual and redundant processes;
- Reduced overlapping stove pipe decision support applications;
- Enhanced customer relations;
- Increased accessibility and more direct access to data for the user(s);
- Consolidated financial data that is more accurate and timely;
- Integrated demographic information and borrower information that can be utilized to determine target areas of greatest need;
- Reduced development costs by producing operations and management reports;
- Decreased need for contractor IT services.

Section 3 USDA Data Warehousing Strategy & Implementation Plan

As described in Section 5, the alternatives for implementing a data warehousing strategy are varied. The best approach for any enterprise depends upon a number of factors, including the size of the organization, political support for the initiative, time constraints and specific business information needs. All these factors were thoroughly evaluated and used to determine the best data warehousing approach for the USDA partner agencies. This section describes the data warehousing strategy that is most appropriate given the environment and goals of these partner agencies.

3.1 Design Strategy

The approach recommended for the USDA's enterprise data warehousing initiative involves the development and implementation of a series of subject-area-oriented data marts, which are commonly linked with common standards. Each data mart will provide an additional piece of the enterprise data warehouse. This approach will ultimately provide a business view of services delivered across partner agencies, and provides more immediate accessibility of consolidated information supporting key business areas, while also laying the foundation for an integrated, enterprise-level data warehouse architecture.

Approach :

- The data marts will be business driven and supported by an Executive Sponsor;
- An investment will be made in the technical infrastructure to support this enterprise-wide implementation;
- The data marts will be built over a period of time, to reduce the overall risk of building one large centralized data warehouse;
- These data marts will be designed and implemented one by one, adhering to the specified interfaces;
- The data marts will be built based upon an enterprise architecture.

Initial data marts will be developed for key business areas within each of the partner agencies. Pilot projects are already underway to move towards this goal. These data marts will be customized for particular groups and linked in a common format. As previously stated, building these data marts based on an enterprise-wide foundation is paramount to later integration efforts.

The majority of business users will perform analysis and reporting against the data mart(s) specific to the business areas in which they work. However, some

business users require analysis and reporting beyond the scope of a single data mart. Reports may require data from multiple data marts for a consolidated view. For example, the report may require loan trends and demographic data nationwide to illuminate potential bias in loan programs when contrasted with similar data of a state or region. The OLAP tools and web portal single point of entry manage the accesses and present an integrated report.

Program managers, agency heads, and above may take a broader view of program delivery across the partner agencies. The data warehouse environment will provide summarized and aggregated data from detailed data marts to meet this need.

Figure 3-1 provides a conceptual overview for the data warehouse-data mart approach recommended.

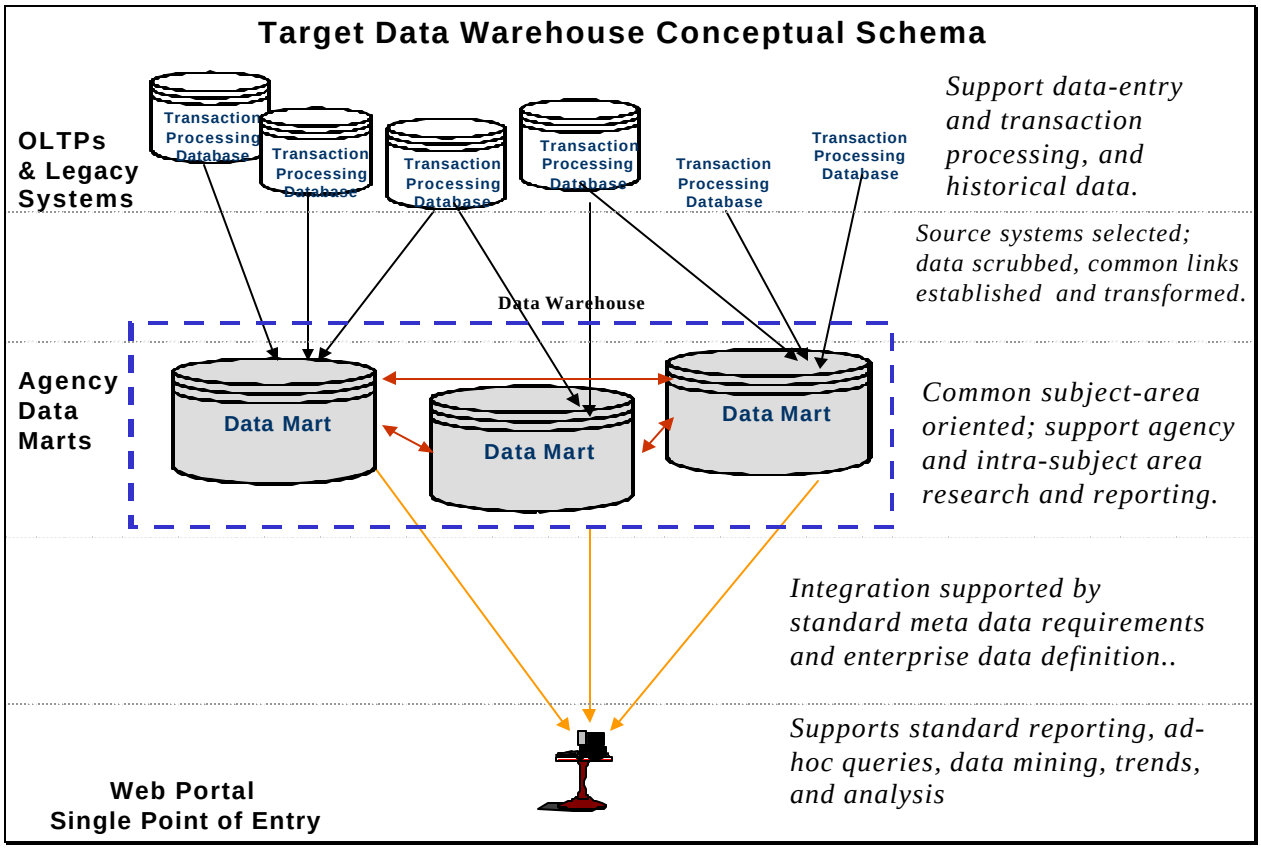


Figure 3-1, Target Data Warehouse Conceptual Schema

3.2 Information Delivery Strategy

As USDA looks at the enterprise delivery of information, it has to be done in the context of how we envision knowledge being presented to our end users. While our data warehouse structure consists of multiple data marts, delivery of information from these data marts is through the primary corporate portal. A portal is the primary point of entry through the web for corporate information.

Concept of a single portal from which all users gain access to information within USDA is depicted in Figure 3-2.

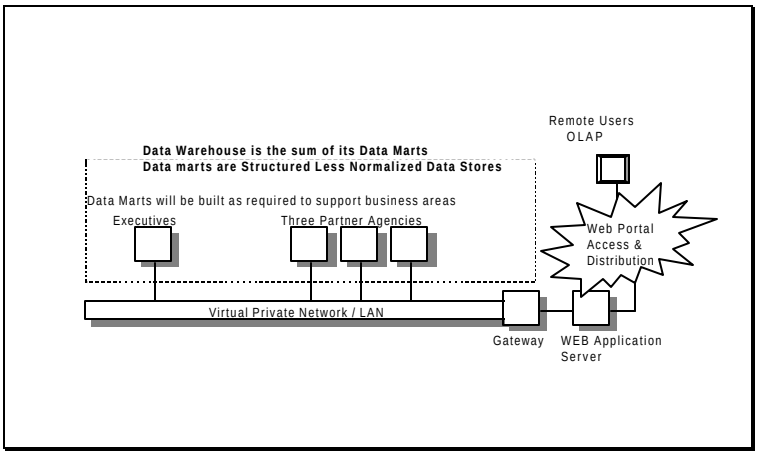


Figure 3-2, Portal Access to Partner Agency Information

3.3 Phased Implementation Plan

The Data Warehousing team will move towards the target enterprise data warehousing architecture on an incremental basis, using a phased development and implementation plan. The enterprise DW implementation phases are depicted in Figure 3-3. The first step in this plan began with the definition of five pilot projects, which are already underway. Valuable hands-on experience gained through these pilot projects will be leveraged in each of the subsequent phases to ensure success. This section outlines the general plan for developing and implementing the data marts using this phased approach.

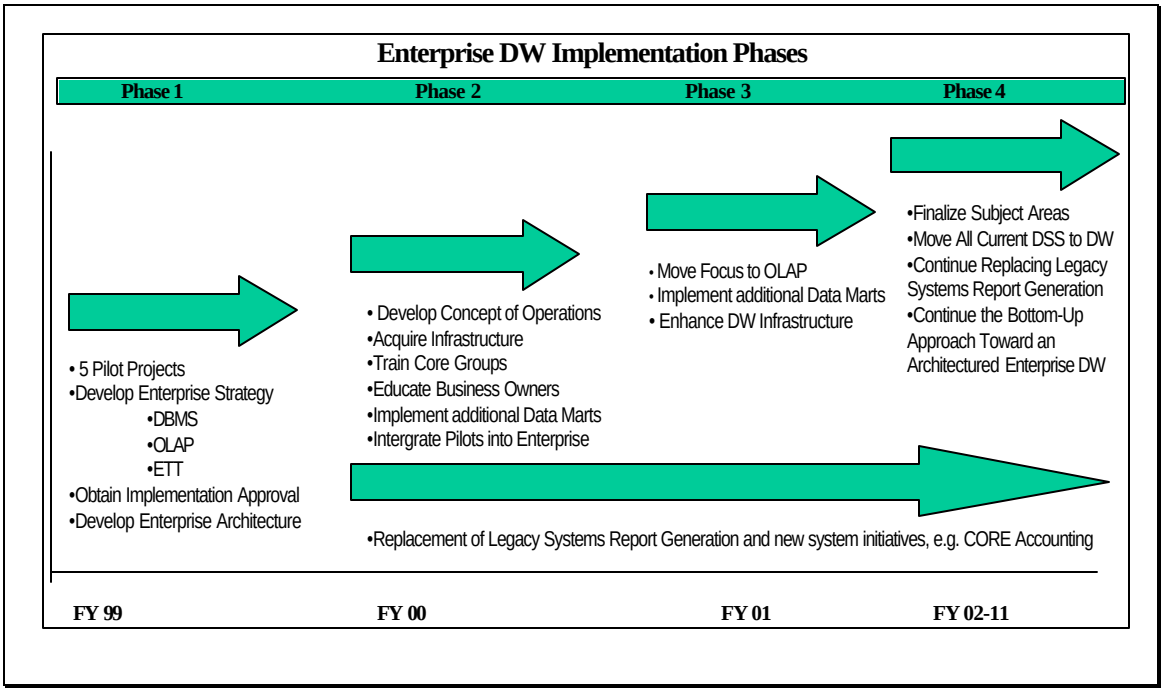


Figure 3-3, Enterprise Data Warehouse Implementation Phases

3.3.1 Phase I: FY99 (Synopsis)

During Phase I, which is currently underway, the Enterprise Data Warehousing initiative was supported by one of the Service Center Data Management Technical Working Groups (TWGs). During this phase, the foundation was laid for the implementation of an enterprise warehouse through proof-of-concept pilot projects.

The DW TWG worked with the Service Center Data Management team to identify the minimum metadata required to support the overall metadata repository initiative from a data warehousing perspective. These teams will continue to work together to ensure that the metadata repository is expanded to account for the DW metadata schema.

The DW TWG also worked with the CCE team to evaluate the scope of the technical architecture being developed to support the infrastructure for the three development centers. The team provided input to the SCI Data Management Tool Selection TWG to support the selection of enterprise data management tools (i.e., relational data base management systems (RDBMS(s)). Input was also provided regarding the enterprise data servers that would assist in determining the server operating system and platform to support an enterprise delivery of data warehousing applications.

During this phase, it was determined that the CCE technical architecture document will not address two components that a successful enterprise wide DW initiative required:

- An Extraction, Load and Transformation tool, and;
- agreement on enterprise on-line Application Processing (OLAP) tools needed to support the end user access to the information.

The DW branch will be responsible for developing the evaluation criteria for the aforementioned tools and will work with CCE for the procurement of the tools.

Phase I affirmed the belief that a successful integration approach to data warehousing requires a “top down” architecture, built from the “bottom up” through architected incremental data marts. These data marts must be business driven and supported by an Executive Sponsor. An investment must be made in the infrastructure to support an enterprise-wide implementation.

3.3.2 Phase II: FY00

During Phase II, the Data Warehousing Branch, which will be part of the Data Management organization under the Information Technology Support Services

Bureau (ITSSB), will take over the implementation of all data warehousing initiatives, including Service Center Implementation Business Process Re-engineering pilot efforts.

Initially, a Concept of Operations will be developed which will outline how the organization plans to implement the data warehousing strategy. The Concept of Operations will outline the other major tasks associated with Phase II; these include addressing the:

- roles and relationships between the organization and the three development centers;
- requisite skills and training necessary for current organizations within the development centers to build core expertise in data warehousing, and how that core team will assist the development teams in incorporating data warehousing techniques during the reengineering and redesign of legacy systems;
- manner in which the DW Branch will perform outreach to the Executive sponsors responsible for legacy systems on the advantages of data warehousing as they plan and budget for the reengineering of those legacy systems;
- method/procedure the organization will utilize to work with the Data Administration branch to;
  - implement the DW metadata schema using the enterprise repository;
  - assist in implementing the Data Stewardship program, train data analysts/modelers within the three partner agency organizations on data warehouse-specific modeling techniques and issues;
  - maintain an enterprise data warehouse data model;
  - review all data-related standards to ensure that they reflect data warehousing requirements;
- implementation of a technical architecture required to support the enterprise data warehouse;
- coordination with the ITSSB organization responsible for legacy system migration to build and refine the timeline for populating the common subject areas within the enterprise model;
- formation of alliances with strategic commercial vendors to provide enterprise-wide implementation support. The branch will become an

active member of user groups and influence direction of vendor product lines;

- data warehousing branch functions which will ensure that each DW initiative follows a standard template that outlines the scope, goals, objective, and the business benefits for each initiative. A detail project plan will be provided for periodic review. The objectives will be quantifiable with performance measures established to ensure that the initiatives meet their overall goals. The objectives will be developed from business driven functional requirements.

During Phase II, a multi-year budget will be developed for the remaining phases of the data warehouse implementation plan.

3.3.3 Phase III – FY 01

The DW Branch will be responsible for implementing the Concept of Operations over the time frame established for legacy system migration. The normal maturation of a data warehouse initiative is to initially focus on query and reporting requirements for differing levels of analysts and managers to include Executive Information Systems.

The next phase in the maturation process is the development of Decision Support System technology, using a variety of techniques for data mining. The DW Branch personnel will interface with an advanced technology group that will investigate the best methods for leveraging the investment the partner agencies have made in providing access to their data. It is projected that the migration of legacy systems will begin during this time frame. The DW Branch will ensure that as legacy systems are reengineered data warehousing techniques will be leveraged to integrate decision support data into the overall DW architecture.

3.3.4 Phase IV: FY 02-11

Data warehousing is a journey and not a destination. It is projected that during this timeframe all major common data subject areas will be integrated into a virtual enterprise-wide data warehouse. The majority of the current legacy systems are targeted for completion of their reengineering/redesign initiatives. During this phase the DW Branch will concentrate on future direction in line with new business requirements and emerging technologies.

Section 4 Data Warehouse Infrastructure Requirements

The successful development and implementation of a large and diverse enterprise data warehouse requires an intricate infrastructure of inter-related technology (hardware and software), data management policies, practices and standards, and staff skills. This section describes the technical, data management and staffing infrastructure necessary to adequately support the USDA data warehousing initiative.

4.1 Technical Infrastructure Requirements & Criteria

The technical infrastructure (hardware, software, communications and security) is a critical element in the data warehousing initiative. Of the technical infrastructure requirements, many are standard (already addressed through the CCE technical architecture), while others are unique to the Data Warehousing initiative.

4.1.1 Standard Technical Infrastructure Requirements

The following technical infrastructure requirements are standard to the CCE.

4.1.1.1 Telecommunication LAN/WAN/VOICE Requirements

The LAN/WAN/Voice initiative will provide the connectivity from the Service Center to the State, and up to the Agency. This is essential to the success of the data warehousing initiative. The data warehousing team will provide an estimate of telecommunications requirements.

An assumption has been made that the LAN/WAN/Voice initiative will provide 56k frame relay to the Service Centers within the 12-18 months of this writing (June 1999). To ensure timely access to critical business information 56k frame relay is a minimum requirement.

The DW team is working with the CCE initiative to provide estimated load requirements to simulate and forecast DW requirements for both the hardware and telecommunications. It has been determined that future DW applications will use browser technology over the Internet.

4.1.1.2 End User Hardware Requirements

The CCE will provide USDA employees with the following configuration, which is more than adequate to support data warehousing requirements:

- Compaq Deskpro EP;
- 400 Mhz Pentium II;
- 64 MB RAM 6.4 GB HD;

- 32X CD-ROM;
- 17" SVGA Monitor;
- Matrox 8MB MGA-G200 AGP Video;
- Keyboard & Mouse;
- 2 Serial Ports, 1 Parallel port;
- 2 USB ports;
- Sound Card with internal speakers;
- 10/100 Ethernet Card;
- MS NT Workstation 4.0;
- 3 Year On-Site Parts;
- Labor Warranty.

The assumption is that the requisite telecommunication servers to allow access to the Internet will be in place by end of FY 2000.

4.1.1.3 Security Requirements

Security concerns have heightened in recent years. News events about computer-related data errors, thefts, burglaries, fires, and sabotage pervade the news. The nature of the computing environment has changed significantly. The increased use of networked computers, including the Internet, Intranet, and Extranet, has had a profound effect on computer security. The greatest advantage of remote access via networks is convenience. This convenience makes the system more vulnerable to loss. As the number of points from which the computer can be accessed increases, so does the threats of attack. More caution is clearly needed to counter such threats. With the advent of data warehousing and the need to provide end-users maximum accessibility, security precautionary measures must be established that will safeguard USDA's data assets. Toward that end, Enterprise Data Warehousing initiative will adhere to the Service Center Security Plan dated November 2, 1998. The plan establishes guidelines for security measures used in the USDA Business Integration System. These guidelines will be properly applied to ensure maximum protection and safeguards for information and resources of the USDA Business Integration System.

4.1.2 Data Warehousing Technical Infrastructure Requirements (Non-Standard to CCE)

The Data Warehousing team working with the CCE has identified two major areas that are not currently covered in the CCE technical architecture. This section addresses those areas; selection criteria for the standard set of Extraction/Transformation/Translate tools and the end-user decision-support OLAP suite of tools are outlined in Appendix A.

4.1.2.1 Data Warehouse Server & Operating System Requirements

The USDA concept for data warehousing shares a feature common to all current data warehousing initiatives, i.e. the full size and scope of the data and metadata are not known, because the policy to determine the amount of history data and policy to define the metadata are not yet defined. Growth also occurs, if the warehouse is successful, because agencies, other than the three partner agencies, will want to participate and increase the scope of the data to include data unique to their part of the enterprise.

Therefore, the hardware selected to house the repository for the warehouse must be scalable to grow with the size of the data and metadata repositories as the growth occurs. As the cost of hardware decreases and as the capability of technology increases, incremental upgrades reduce the total cost of ownership. According to the Gartner Group and Meta Group the greatest reductions in total cost of system ownership is achieved in consolidation of information and administration supported by a data warehouse.

Specific criteria for the USDA data warehouse initiative include:

- For performance and efficiency, the architecture of the hardware and its operating system must provide for multi-threading and parallel applications;
- Multi-threading and parallel processing hardware features are useless if the DBMS software does not effectively exploit these features;
- The operating system must provide dynamic and dedicated processor resource allocation mechanisms to achieve high computer resource use factors;
- The hardware and its operating system must be based on public and generally recognized standards, so that their products properly function in an heterogeneous multi-vendor environment;
- The vendor of the hardware and its operating system must have a demonstrated successful track record in supporting its products;

- The vendor must have a commitment to improve their products by using the best and recent technology in the design of their products to continue to reduce the total cost of ownership;

- The vendor must have an independent assessment (benchmark) of the quality of their products in warehouse use, and the assessment must receive high ratings with peers. For example, the Transaction Processing Council may have executed their standard test series D (data warehouse uses) with high scores for performance and stability;

- The vendor must be stable in its market place to enhance the likelihood that they continue to do business in the marketplace. The vendor must make provisions available to the USDA so that source code and other proprietary materials are available to the USDA in the event that the vendor no longer performs in the market during the life of the system.

4.1.2.2 Relational Data Base Management System (RDBMS)

The Data Warehousing team developed a document that is intended to act as supplement to the SCIT Data Management Tools Selection Strategy document for the express purpose of articulating the baseline requirements for the Data Warehouse Initiatives. A Data Warehouse Technical team has been established for the express purpose of addressing this subject issue and is to be the point of contact for coordination by the CCE for test evaluations and selections.

DBMS products will be evaluated for their support of each category of application. Different DBMS products may be considered to meet different application requirements, but the evaluation will give stronger weighting to products that can support multiple application categories. This is because fewer products will result in less training, administration and lower cost of ownership.

The DW team submitted the evaluation and criteria to CCE for selecting DW Enterprise RDBMS(s) on June 10, 1999.

4.1.2.3 Extract/Transformation/Translate (ETT) Tools

Data extraction, transformation and translation tools (ETT's) extract data from multiple data sources and, using business rules, massage and transform it into new information in preparation for loading a target database. These tools are used for a wide variety of data movement needs, including data warehouses, data marts and application or database conversions. Appendix A outlines the evaluation criteria for selecting an ETT tool.

4.1.2.4 On-Line Analytical Processing (OLAP) Tools

In addition to the RDBMS and ETT tools used by the data warehouse development and maintenance personnel, another suite of tools is necessary to put reporting and data-mining capabilities into the hands of business users. OLAP tools will be selected as part of the Data Management tool set. Appendix B outlines the evaluation criteria for selecting OLAP tools.

4.2 Data Management Infrastructure

A cornerstone of any successful data warehousing initiative is a solid data management/data administration program. A data warehouse is essentially a tool to facilitate integration and sharing of business information across disparate organizations and databases. Without the implementation of critical data management principles, such as standards for common data and metadata, these goals are not possible.

In addition, important data management infrastructure components such as the enterprise data model and the repository provide the means for identifying common and shared data needs, prioritizing critical areas of data for initial focus, and identifying existing databases and applications that may provide source data for the warehouse environment.

The data management infrastructure components which are critical to the success of the Data Warehousing initiative include:

- the enterprise data model;
- standards for data naming and definition;
- minimum metadata requirements and standards for data warehousing.

The enterprise data model provides an over-arching context for the information requirements that span the agency, from the Service Center perspective. It is used to define high-level data subject areas and to prioritize those subject areas for data mart focus. The initial focus will be on areas of common data and shared data, where return on investment is expected to be greatest. Figure 4-1 illustrates the concept of common data, shared data and unique data; the definitions for each of these terms following the diagram.

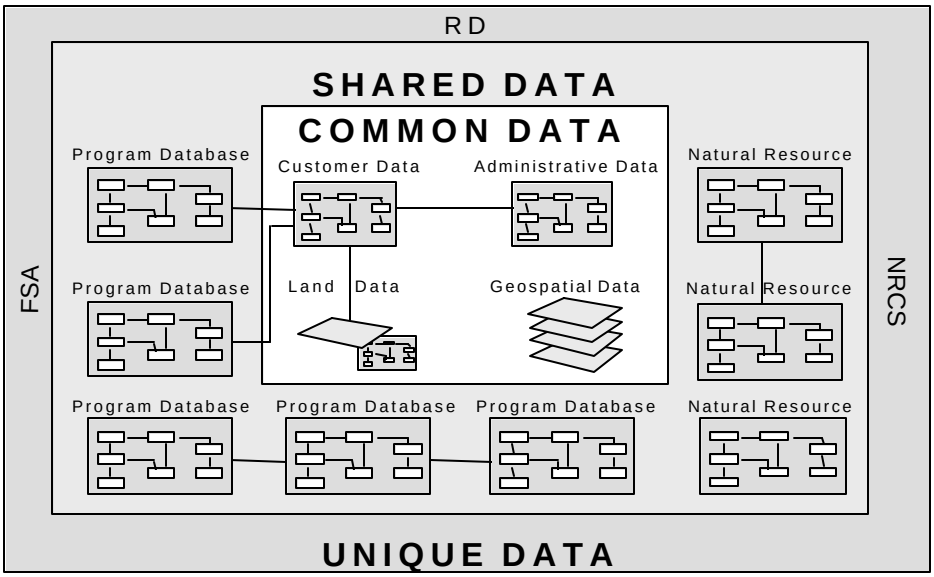


Figure 4-1, Service Center Enterprise Data

***Common data** is data jointly owned, used, and managed by Service Center partners. The common areas of interest for the BPR initiative are common customer data, office data, administrative data, common land unit data, and standard geospatial data.*

***Shared data** is data owned and managed by a specific Service Center partner and shared by other partners. The key areas of shared data include the loan, and financial (debt information) data.*

***Unique data** is data owned and managed by a specific Service Center partner but not shared across agencies.*

Standards for naming and defining data entities and attributes (including domains, formats, business rules, etc.) and identifying standards keys are also defined and managed through the enterprise data model. These standards are of paramount importance to the success of a data warehouse. Without standards it is nearly impossible to identify like data or distinguish amongst unlike data in source databases, or to adequate define the transformation rules as that data is mapped into the data warehouse. Without intelligible business names and robust definitions, users of the data warehouse cannot correctly identify and therefore properly use the information.

One of the goals of the enterprise data model is to identify standard, globally-unique keys for each data entity (including tabular and geospatial data) defined as common or shared need. These keys, which would be carried through the data warehouses as well as OLTP systems, would enable users to access data

seamlessly across servers, laptops, service center computers, offices and even over the Internet to customers. Consider for example, the situation where each agency might utilize a different piece of information to uniquely identify a customer (such as Name, Customer-ID Number) across the different systems, and how much easier it would be from the end-user and customer perspective if one key were used across all systems.

The Data Warehouse initiative will also be a critical participant in the on-going cycle of metadata collection (registration) and maintenance. Metadata items, which must be registered in the repository, include data elements groups, data elements, domains, data sets and software/application systems. The on-going maintenance of this information will ensure that it remains a valuable resource for continuing data management and development efforts as the concepts of data sharing, warehousing and integrated business intelligence move towards maturity.

4.3 Staff, Skills, Positions & Training Requirements

The Data Warehousing Branch, Data Management Division, Information Technology (IT)/Support Services Bureau (SSB) will be responsible for technical support to the partner agencies' customers, including the various program offices. There will be approximately 12 full-time equivalent (FTE) staff positions devoted to data warehousing support. The skills of the personnel assigned to the Data Warehouse Branch will vary from novice to senior-level data warehousing experience. The pilot projects offered invaluable experience, to partner agency personnel that will be supplemented with formal training where appropriate.

Immediately following creation of the SSB, a skills assessment among members of the Data Warehouse Branch will be performed. The following types of skills are necessary. If these skills are not deemed to be available through Data Warehouse Branch staff, additional training will be required:

- **Data Warehouse Concepts & Facility** - Understand the basic concepts, components, and facilities of a data warehouse and the business benefits of a data warehouse approach to information management and delivery;

- **DW Architecture and Infrastructure** - Understand the characteristics and the basic technical framework of a data warehouse and the various infrastructure components that can be implemented under a chosen architecture;

- **Business Process Re-Engineering** - Ability to develop business process models and incorporate new business practices and technologies. The experience will focus on implementation,

continuous innovation and organizational re-engineering for best practice;

- **Data Analysis & Modeling** - - Experience in analyzing and modeling data relationships using Entity Relationship (ER) diagrams and, presenting the results within the framework of the business requirements. Ability to understand business rules and practices, as they pertain to data, and communicate those rules through data models, the data dictionary and other supporting components;

- **Performance Tuning -** Experience in conducting performance tuning in database and applications;

- **Data Propagation Tool Experience** - Experience in using data propagation tools in extracting, cleansing, transforming, aggregating, summarizing, indexing, and mapping source to target DW database;

- **Query/Report Tool Experience** - Experience in using various query/reporting tools to maximize or reinforce the data delivery function in a data warehouse environment;

- **Marketing & Communications Skill** - Ability to communicate effectively with the user community on data warehouse activities and business payoffs, spend time internally selling the data warehouse to user departments, develop a strategy to encourage use of the data warehouse facilities, and excite interest in using the data warehouse while simultaneously managing expectations.

It is envisioned that there will be at least the following basic roles required to effect a solid, comprehensive, and successful Data Warehousing initiative. It's possible for one individual to play multiple roles or for one role to be played by more than are individual:

- **The Data Administration (DA)** role includes the people responsible for defining better ways to identify, represent, organize, store, and provide data to effectively satisfy the information requirements of business users. This is accomplished by designing and developing the models and logical database structures on which the physical structure will be developed. They will also manage the metadata that is critical to end-user success in using the data within the DW;

- **The Database Administration (DBA)** role includes physical design, development, and management of databases for the DW. The essential responsibilities to be performed by the DBA is review of the logical data model, design development, and implementation of physical database structures, database, utility functions, and database tuning;

- **Business Analysts or Decision Support Analysts** understand the business processes and the data requirements in terms of data entity, attributes, and relationships, and are responsible for working with information users to meet requirements in the areas of data usage and access. These people may also assist in the development of decision support applications. In addition, this group provides analysis support for development and execution of business plans, micro-management strategies, and other day-to-day activities in support of the business objectives;

- **Application Planners and Developers (AD)** determine the application architecture jointly with other support groups and are responsible for designing, developing or acquiring, testing, maintaining, and managing the business decision support applications. The major responsibilities to be performed by the AD resources include Application Project Management, Development, Testing and Evaluation, Client Tool Definition and Selection;

- **Data Warehouse Administrators (DWA)** plan, measure, coordinate, implement, and support the data warehouse environment, data acquisition, data delivery, and associated tools. The DWA group is accountable for the business value of a specified collection of data, prioritization of the development of DWA resources, user satisfaction, DW Marketing, business application opportunities, and return on investment (ROI)

Other support such as Infrastructure, operations and telecommunications, and data security will be provided by personnel belonging to other organizations within the IT/SSB, but outside of the Data Management Division.

Personnel from the Data Warehouse Branch, working with the appropriate IT/SSB customer relations liaison person that interfaces with the various program areas, will determine training needs for the information or business user. The business user could be a functional user, knowledge worker, decision maker, business executive sponsor or anyone having a need to access the Data Warehouse. Business users will range from novices to power users, and training requirements will vary accordingly.

Users will need appropriate training in the use of front-end tools, use of the data directly; use of data models and metadata; development and use of ad hoc and canned reports, queries, spreadsheets; and use of custom applications. Last, but not least, they must be given an understanding of the actual data structures that they will be using on a regular basis. Experienced warehouse developers believe that the average user cannot initially understand more that about 20 tables and their inter-relationships. Users should be given a manageable number of tables and training in the use of these tables. It is not enough to give users training in the front-end tools and use of the data directory.

Section 5 Data Warehouse Terminology & Concepts

This section provides a description and discussion of data warehouses, data marts and decision support systems, as well as strategies for development. These concepts are the major components of a data warehouse centered information infrastructure, and a common understanding of what they are (and are not) is essential to the requirements, development strategies and implementation plans described in the remainder of this document.

The following descriptions are common industry definitions, and are therefore generic. Strategies and plans for how these components might best be utilized within the USDA data management framework are discussed in the later sections of this document.

5.1 Definitions

5.1.1 Data Warehouse

Data warehousing is the coordinated, architected, and periodic copying of data from various sources into an environment that is optimized for analytical and informational processing. The environment is called a data warehouse, in which enterprise-wide information is cleansed, integrated and organized. The warehouse contains a nearly current state and history of the enterprise, which is subject-oriented and time-variant, and is designed to hold large amounts of data. The warehouse is the union of its component data marts.

The true value in the data warehouse lies in its ability to integrate key pieces of data from disparate OLTP systems. Data warehouses do not support transaction processing, and are thus structured differently then databases designed for that purpose. The data warehouse provides users with a clearinghouse from which to combine and crosswalk information from different subject areas, geographic regions, etc. This provides powerful research, trend analysis capabilities and supports executive decision making at the highest levels.

A typical data warehouse may contain a billion bytes of data in different databases with different formats. To be able to transfer data from one database to another so that all data elements are integrated within the data warehouse, programs are needed to reformat, integrate, and transfer data from one database management system to another. As a result, end-users can request data from the data warehouse without the knowledge of the data's type or format. Since the purpose of the data warehouse is to make organizational data more available, the warehouse must include tools not only to deliver the data to end-users, but also to transform the data for decision makers for analysis, reporting, and query.

5.1.2 Data Marts

A data mart is a logical subset of the complete data warehouse. A data mart serves a business unit and contains the data dimensions necessary to support the business unit's mission and standard processes. All data marts must be built from conformed dimensions and conformed facts and employs a set of definitions common to all elements of the warehouse.

Data marts are often subsets of a larger data warehouse. They may be separated from the larger physical data warehouse or virtual data warehouse concept based on the need to meet specific or localized business needs, or because of the development strategy employed.

The information contained in data marts may be refreshed from either the legacy systems or from the source/OLTP systems using extract and transformation tools. Data marts are components of an overall data warehouse architecture.

5.1.3 Data Mart Development Components

Once the business need for a data mart has been identified and agreed upon, there are several "standards-based" components that must be included in a project plan. These standards control the interfaces among components and their behavior. Like more conventional data base applications, the initial steps involve identifying the information requirements and structuring those requirements into a workable data model. Like traditional data bases, a complete data dictionary, robust definitions and clearly stated business rules are required to support this data model. This ensures that all instances of a data entity conforms to the same set of standards. Business subject area data stewards, experts and technical personnel validate the information.

From this point forward, components of a data warehousing project are more unique. The design and construction of a data warehouse involves the following steps:

- Identification of Source Data;
- Data Extraction;
- Data Cleansing or "Scrubbing";
- Transformation;
- Loading;
- Indexing;
- Aggregation;

- Replication/Synchronization;
- Access & Analysis;
- Data Integrity Testing;
- Distribution;
- Data Administration;
- Training;
- Security and Integrity.

5.1.3.1 Source Data:

Traditionally the warehouse starts with identification of source data against an approved and authoritative data model. The bulk of source data comes from internal OLTP (On-Line Transaction Processing) applications and external legacy operational systems. One of the underlying assumptions is that an extract will be obtained from the OLTP system for current updates and legacy systems for history. The three farm agencies have a tremendous investment in data on the NITC mainframe in KC. The mainframe acts as a data store, as well as provides OLTP type of applications functionality. For the most part the access to the legacy information is provided via reports. Traditionally those reports have been hard coded mainframe reports that require contractor support in modifying and maintaining. A data warehouse or mart would provide users with easy access to agency information.

5.1.3.2 Extraction, Cleansing, Integration and Transformation Process:

The most time consuming and costly process is that associated with the extraction of the data from various sources, the cleansing of the data to adjust for abnormalities and creating the ability to relate the information from disparate applications. Extraction, transformation, and load tools will play a pivotal role in automating the management of a data warehouse. These tools will integrate the source and target systems, providing all necessary transformations on the data mart prior to loading it into the data warehouse or data mart. For example, there are at least four different ways USDA applications address gender code. These alternative formats are transformed to a single internal standard definition. In order to put a process in place that provides for recurring extraction, cleansing and transformation, careful planning is required.

5.1.3.3 On-Line Analytical Processing (OLAP) Tools

On-Line Analytical Processing (OLAP) tools enable users to quickly analyze information that has been summarized into multi-dimensional views and

hierarchies, such as that in most data warehouses and data marts. They can be used to perform trend analysis, or to drill-down into masses of transaction history in order isolate particular pieces of data.

OLAP tools perform all the functions previously performed by decision support applications. These OLAP tools permit web enabled access to the data. The user defines the nature of the analysis and/or report and the OLAP tool via the web portal, locates the information from any of the data marts and returns an integrated report, published to the web for the user. A portal is the primary point entry for corporate information.

5.1.3.4 Meta Data

Historically, the often-coined definition of meta data is "data about data" - basically, information about the information architecture within an organization or within an enterprise. But meta data can be much more. A meta data store could include end-user security profiles or data dictionaries describing tables, data types, attributes within tables, join conditions and transaction data. It might hold transformation rules that describe how to move data from a transactional database to a multidimensional target database, and the aggregations and calculations that are applied to it. It could also hold common business rules (such as definition of a loan or borrower) that multiple tools may want to access.

Meta data also allows end-users to move easily from one data source to another. Sometimes meta data reveals that information to the end-user, but often the tool moves from one data source to another transparently, without the end-user ever knowing it has happened.

The typical data warehousing implementation is rarely a single transaction database with a single transformation, scrubbing or data movement tool, or a single homogeneous data warehouse with a single type of business intelligence end-user platform. It's a lot messier than that. There are multiple boxes all over the place, multiple forms of source data, and multiple ways of moving data from transaction systems into a store that's tuned, organized, and scaled to accommodate querying, reporting, and OLAP analysis.

There can be multiple islands of data warehouses. These may include dependent data marts where someone has organized corporate-level information into a data warehouse and then spawned smaller data marts off it. More often than not, however, business pressures and the need for rapid Return-on-Investment (ROI) spawn, independent data marts that feed off Department-specific transactional databases that have no relation to a central enterprise source.

This presents real problems for administrators who must create an infrastructure that makes both transactional and historical data available for analysis for users and manage it using front-end business intelligence tools. Meta data is the key to successful management and the administration of all this data. If administrators

are going to maintain any level of sanity in achieving fast and reliable access to information, they need a strong meta data store.

5.2 - General Data Warehousing Strategies

There are a number of varied approaches to designing, building and fielding an enterprise data warehouse. At opposite ends of the spectrum are designing and building a "top down" data warehouse; the other extreme is designing an enterprise solution but building from the "bottom up".

The "top down" approach mandates the, construction of an enterprise data warehouse first, then the distribution of subset data marts from that parent data warehouse[2]. The "bottom up" approach calls for the development of a series of incremental, architected data marts with the end-goal of integrating them into an enterprise data warehouse.

Although the "top down" strategy was favored in early enterprise data warehouse projects, and is the considered to be the most elegant design approach, high rates of failure for initial enterprise data warehouse projects have led the majority of current projects to the "bottom up" approach. Each strategy has inherent strengths and weaknesses, and should only be used where appropriate. Although each is viable under suitable circumstances, the misapplication of strategy can pose considerable obstacles for a project from the start.

When assessing the appropriate approach for individual circumstances, it is important to have an understanding of the benefits and drawbacks of each approach.

5.2.1 "Top-Down" Approach

The benefits of a "top down" strategy include:

- A "top down" strategy delivers a data warehouse that **is inherently architected.** Every subset data mart that is created from that parent data warehouse stands to inherit this architecture. This removes or vastly minimizes the effort to integrate across the data marts, and eases the maintenance burden;

- It provides **enterprise perspective.** It supports viewing, summarizing and analyzing data from the activity level up to the enterprise level;

- It provides a **central metadata repository** for the system, making the maintenance of the system much less complex than it would be with multiple metadata repositories;

[2] The Intelligent Enterprise, Richard Tanler, Bill Inmon, Douglas Hackney, Datamation

- It supports **centralized rules and control** of the data warehouse. The "top down" approach ensure that there is one and only one set of data extraction, cleansing, and integration jobs or processes to monitor and maintain, thus minimizing complicated data replication and synchronization issues.

Despite these attractive benefits, there are also several drawbacks pertaining to the "top down" approach. These include:

- **The time required to develop and implement an enterprise data warehouse is considerable.** Enterprise data warehouses are typically built in an iterative manner, subject area by subject area (such as customers, finance, human resources, etc.). Depending on the size of the enterprise involved, it may typically require three to four years to implement a data warehouse. This is a long time to maintain political and budget support in the face of ever-shifting priorities, emergencies and team members. Further complicating the building of an enterprise data warehouse is that the effort does not have clear directions from the business side of the organization. As a result, the project can take on a life of its own without any end goal in mind. When the scope of the project is too big and without any well-defined results, it can continue forever, with no tangible results;

- There is a **high exposure to risk**. Enterprise data warehouses are inherently high exposure ventures because one of the prerequisites for success is the sponsorship, and therefore the attention, of the Chief Executive Officer (CEO) and Board of Directors (BOD). At this level of exposure, there is little to no room for error;

- Top down" enterprise data warehouse development requires a high-level of cross-functional management skills and support.

5.2.2 "Bottom Up" Approach

The benefits of pursuing a "bottom up" approach to data warehousing include:

- **Implementation is considerably faster** than with a "top down" approach. Data marts are essentially mini-data warehouses without the huge cost, long time investment, and high risk of failure. They are ideal for a rapid, iterative, prototype deployment. Data marts should not be used as a cheaper solution for a data warehouse; they should represent an initial step toward an enterprise data warehouse. Data marts should be designed to integrate with a future enterprise data warehouse, or much rework will have to be done over the long term. Because incremental architected data marts are built to tightly focus on a specific business area, they can often be brought to production

status in six to nine months (depending on the scope and size of each data mart).

- This approach offers a **faster return on investment (ROI).** It is very challenging to maintain political will in the organization for the typical 9+ months required to deliver an initial subject area of an enterprise data warehouse. An incremental architected data mart can demonstrate value and return to the business much faster, and provide a foundation for further investment by the business with a higher level of confidence in future/follow-on efforts;

- The more focused scope of the individual data marts created through the "bottom up" approach minimizes one of the greatest management challenges in data warehousing, which is **keeping the team focused** on a deliverable scope. Incremental architected data marts are inherently focused on a specific business area, thus limiting the trend of projects expanding to areas and teams expending efforts on issues that are not directly related to the business area being addressed;

- The approach of building individual data marts before integration at the enterprise level is **inherently incremental**. This strategy mandates a step-by-step approach to delivering information assets. In addition to speeding ROI, it also minimizes risk exposure by allowing the team to grow and benefit from lessons-learned at each step. Tools, technologies, consultants and vendors can be subjected to a "test and toss" process. Those that work can be retained for the next step, those that don't can be replaced, with a natural breaking point between incremental steps.

There are also drawbacks to the "bottom up" approach to data warehousing. Some of these include:

- One of the more serious dangers of this approach is the creation of non-architected ("stovepipe") data marts that cannot be adequately integrated to support the enterprise view of data. The advent of easy-to-use drag-and-drop tools has facilitated the temptation to simply build an individual solution to an individual business need, with little or no regard for the overall enterprise architecture or perspective. These stovepipe data marts often become legacy data marts, or legamarts, and are difficult or impossible to integrate at a later date;

- It is challenging to obtain the enterprise view through data marts that have been developed incrementally, even when they were built around an enterprise architecture. It requires more time and effort to extract answers at the enterprise level when the data must be extracted from several individual sources and combined than it would from an enterprise data warehouse. While data marts allow for a smaller

hardware investment and localization of data, they cannot provide the organization-wide business analysis or economics of scale provided by a central warehouse server;

- Managing and coordinating multiple data mart initiatives and teams is challenging. Incremental architected data marts are very popular to build in parallel. This can lead to management challenges in trying to coordinate the efforts and resources of multiple teams, especially in the areas of business rules and semantics;

- Incremental architected data marts are often burdened with the "curse of success". In these cases, the target users of the data mart are overwhelmingly happy, while wanting more information added to their data mart. At the same time, other homogenous user groups in the enterprise are clamoring for their own incremental architected data mart. This leads to political, resource and management challenges for the data mart team(s).

5.2.3 Assessing the Appropriate Strategy

To understand what strategy is appropriate for a particular enterprise, consider the following factors. The market may be roughly divided into three general types of organizations:

- Those that "think globally and act globally". These organizations are driven by strategic considerations and do little that is not reflective of this. They represent the bulk of the early adopters of "top down" enterprise data warehouses. Although many (if not most) of them failed in their initial efforts, nearly all kept trying until they successfully built some version of an enterprise data warehouse;

- Those that "think globally and act locally". These organizations think in strategic terms, but choose to execute on a local basis as needs dictate, while keeping their efforts in alignment with enterprise strategic goals. These organizations represent the bulk audience that the data warehouse market is moving into today. This is the strategy recommended by the USDA Data Warehousing Team;

- Those that "think locally and act locally". These enterprises do not have the organizational maturity or the political processes required to articulate and execute a strategic vision or direction. They are marked by tactical efforts undertaken to solve tactical challenges. Because these projects are nearly impossible to integrate at a later date, whatever is built is eventually thrown away.

5.3 Conclusion

To be successful, the appropriate DW strategy for the USDA requires a "top down" architecture, built from the "bottom up" through architected incremental data marts.

This robust top-down architecture is essential to integrate incremental data marts. This integration will be smoother, less resource-intensive, and ultimately more successful when the data marts are built using this foundation.

A data warehouse-data mart architecture provides flexibility and extensibility to support different levels of information needs. This gradual development approach requires fewer development resources and a lower overall investment.

This architected approach will provide the best data standardization/ consistency and data sharing opportunities, as well as the greatest initial return on investment.

Appendix A: Selection Criteria

A.1 Extraction, Cleansing, and Transformation Tools

Data extraction and transformation tools (ETT's) extract data from multiple data sources and, using business rules, massage and transform it into new information in preparation for loading a target database. These tools are used for a wide variety of data movement needs, including data warehouses, data marts and application or database conversions. In addition to the major technical requirements described below, the vendor should be evaluated for financial soundness, and support.

A.2 ETT Implementation Architectures

The initial set of vendors in the market focused on the generation of application code to perform the extraction and transformation of data during the migration process. More recently, vendors have focused on non code-generating technology with less emphasis on source data extraction and have instead focused on providing frameworks for the specific transformation of data, leaving users to develop the extraction process. This process is called transformation engine. Users employ specialized tools or raw coding to extract legacy data and place it into a relational database. Then the transformation engine is used to provide further editing and transformation for the target warehouse database.

Neither architectural approach (code-generation or transformation engine) meets all possible needs. They are complementary rather than competitive.

A.3 Major Evaluation Criteria Categories

Evaluation criteria can be specified within 8 major categories:

- Data Identification and Extraction;
- Record Matching, Merging, and Integration;
- Data Cleansing;
- Data Transformation;
- Metadata Management;
- Ease of Use and Development;
- Operations Management and Database Loading;
- Scalability and Performance;

A.4 Data Identification and Extraction

This process supports the identification and extraction of records that meet user-defined selection criteria. Basic evaluation criteria are:

- Tool should support existing databases;
  - No limitation on the number of data sources that can be extracted simultaneously;
  - Method should support database as well as log-based sources;
  - A changed data capture facility uses log data or triggers (e.g. date/time stamps) to obtain net changes for periodic updating of the data warehouse;
  - The ability to move data back and forth between two different database management systems;
  - Ability to read delimited flat files, spreadsheets.

A.5 Record Matching, Merging and Integration

The ability to match records across multiple sources concurrently facilitates the identification of redundant and overlapping data. Basic evaluation criteria are:

- The ability to match records (either by appending a common key or using data- driven matching or fuzzy logic) and then combines multiple records into a single new record. This helps eliminate duplicate information (e.g. customers, addresses).

A.6 Data Cleansing

These features ensure the quality of the source data. It is the critical phase of the ETT process. Basic evaluation criteria are:

- Check all data types;
- Validate field values and range of values;
- Match data to other code files;
- Place invalid records into a suspense file;
- Substitute default values in field that are incomplete;

While a critical phase of the data warehouse process, this is one of the weak areas of vendor products. A sub-market has evolved to address this functionality.

A.7 Data Transformation

Data transformation refers to applying processing routines to data to transform it into new formats and/or values for a target database. Basic evaluation criteria are:

- Business rule-driven transformations (often using look-up tables);
- Pre-built transformations (such as date or money conversion routines);
- Date type conversions;
- Standard abbreviation conversions (e.g.. convert Street to St.);
- Callable user written routines;
- Value substitutions (01000 = 01003);

A.8 Metadata Management

Metadata is information about the business and technical characteristics of data that that supports the business. Such characteristics are common business name, description of the data, field type (numeric, date), field length, acceptable values and ranges, etc.

In the ETT process, both code-driven, and transformation engines use Metadata Repositories to drive the extraction, cleansing, transformation, and loading process. Evaluation criteria are:

- Repository sharable across applications;
  - Interfaces to application repositories (e.g. PowerDesigner), and computer-aided software engineering tools (CASE) like CoolGen;
  - Supports impact analysis;
  - Notify developers if a change to legacy structures requires the ETT process to be re-executed ;
  - Allows for input by a wide group of users – not just Data Management Specialists;

- Ideally this repository should also support the actual data warehouse. However, this level of integration is not mature. Thus import and export of metadata from one repository to another is desirable.

A.9 Ease of Use and Development

The potential benefits of these tools are increased programmer productivity and program maintainability. Evaluation criteria are:

- Graphical User Interfaces (GUI's);
- Impact Analysis;
- Rapid prototyping and testing;
- Version control;
- Ability to call external routines;
- Ability to integrate other tools;
- Capability to enter external routines into the repository for future use

A.10 Operations Management and Database Loading

ETT processes become on-going tasks, and managing these processes is critical to their success. Evaluation criteria are:

- Built-in audits;
- Checkpoint/restarts;
- Error logs;
- Operational reports;
- Alerts;
- Scheduling function (clock-based, event-driven, triggers);
- Generation of JCL, Unix scripts, etc;
- Load data directly from the tool using Application Program Interfaces (API's);

- Load data using the Relational Database Management Systems (RDMS) load utility. These utilities usually take flat files as their input;
- Parallel loading for large data volumes;
- Ensuring referential integrity during the load process;
- Support for bit mapped indexing, partitioned tables;
- Build aggregations while loading the data.

A.11 Scalability and Performance

Scalability and performance needs to be evaluated:

- Total throughput time;
- In-memory management to minimize disk access;
- Scalable processing components to allow for parallel processing on the same platform or across multiple platforms, (code-generating tools generally operate on a single platform.

Appendix B: On-line Analytical Processing

B-1 Web client

1. The web client must be able to be installed directly onto the client workstation from the server system with no user intervention. Defined as ZERO Administration Clients. Reducing installation, maintenance and version control costs.

2. The web client must be able to be upgraded automatically to new version releases from the server system with no user intervention.

3. The web client product must be easy enough to use for non-technical personnel with limited training time invested.

4. The web client product must provide one single, integrated and easy-to-use tool that includes the ability to launch a query, analyze the data through the use drill-down capability, chart and graph the information and build custom reports.

5. The web client product must be able to be deployed in a read only version and a full analysis version that supports the capability to perform an ad-hoc query, charting, analysis and reporting.

6. Product must allow the user to disconnect from the source database system after they have retrieved the results of a query and work off-line from the database server

7. The product must have the ability to navigate down through information to determine why events may have occurred within the query results, without having to pre determine where a user may need to go.

8. To reduce training costs and time, it is desirable for the web client and the client/server client must have the same look and feel.

B-2 Client/Server Client

1. The client/server client product must be easy enough to use for non-technical personnel with limited training time invested.

2. The client/server client product must provide one single, integrated and easy-to-use tool that includes the ability to launch a query, analyze the data through the use drill-down capability, chart and graph the information and build custom reports.

3. Product must allow the user to disconnect from the source database system after they have retrieved the results of a query and work off-line from the database server

4. The product must have the ability to navigate down through information to determine why events may have occurred within the query results, without having to pre determine where a user may need to go.

B-3 Report distribution

1. The product must have the ability to perform a particular canned analysis report once or on a repeatable scheduled basis and then automatically distribute the report out to the entire user community.

2. The product must offer an easy mechanism to deliver our standard reports to all users within the USDA Intranet

3. At a minimum the product must be able to deliver these reports to the users via: Email, File Server, Web Server, Print Server and to HTML pages

4. The product must offer an easy mechanism to deliver our standard reports to our POCs and other users.

5. At a minimum the product must be able to deliver these reports to the users via: Email, File Server, Web Server, Print Server and to HTML pages

6. The pre-planned analysis reports that are distributed to a wide audience must have the ability to suppress functionality in the report such as the capability to drill-down or the capability to perform a new query on that report. The distribution capability for any given report must provide a read only capability, a full analytical capability and/or an ability to perform a new query.

B-4 Query

1. The product must allow the ability to perform a heterogeneous data base join between two separate database systems.

2. The product must not allow users to edit database fields, i.e. a “read only” query tool.

3. The product must allow a user to edit and write custom SQL code.

4. The product must allow connectivity to all major supported relational database systems.

5. The product must allow the ability to import and export query results to Excel and Lotus spreadsheets.

B-5 Security

1. The product must maintain document level security based upon user login and password. The ability to access the database, analyze the provided information, and change the format must be restricted.

2. The client product cannot have any Java code embedded into it due to current security requirements. Must certify.

3. The product must have built in auditing features to evaluate successful query completion, average times to process, and peak activity. This auditing capability must be able to determine data utilization, query duration and volume for database performance tuning.

B-6 MISCELLANEOUS

1. The product must be Year 2000 compliant. Must certify.

2. The product must be able to support all of the hardware platforms throughout USDA including Windows 95, NT, Macintosh and Unix.

3. The product must support our initiatives for rapid prototyping. After the database/data warehouse system is solidified the time to first end-user query should be measured in hours.

4. The product must have an integrated and comprehensive help function.

Appendix C: Data Warehouse Pilot Projects

C.1 Pilot Data Mart Projects

Pilot projects were defined to initiate the data warehousing initiative and to assist in determining the enterprise architecture. A description of three of the representative projects is attached as Appendix D, E, and F.

These pilot projects assisted in validating the benefits of data warehousing for the business community, and provided the opportunity to glean technical information upon which to develop the enterprise strategy. Results from the pilot projects were used to determine the selection criteria for the standard suite of data warehousing tools for the partner agencies.

C.2 Lessons Learned

The pilot projects provided not only a proof of concept for the business community, but also for the data warehousing development team to glean valuable technical information and hands-on experience with data warehousing hardware and software tools. Benchmark performance measurements and other evaluation parameters yielded technical information that was instrumental in determining criteria for the selection of the USDA standard suite of data warehouse tools.

The development and implementation of the pilot projects yielded many valuable lessons. Some of these lessons are:

1. First and foremost, procuring the services of an experienced data warehouse consultant is worthwhile. The data model is the foundation on which the data warehouse is built, and has special requirements above and beyond those required for traditional data base applications. A certain degree of expertise is required in order to design the most effective and efficient model to meet customer needs. The expertise of a specialist in designing enterprise data warehouses is required.

2. The consolidation and summarization of business information in the data marts will likely identify gaps in the data. New software may need to be developed or existing software modified in order to fulfill the need for the data that does not currently exist.

3. The scrutinizing of the data that goes into the data warehouses inevitably identifies issues involving the quality of the data, in terms of both consistency and accuracy. The quantity of this data that must be cleansed is invariably more than anticipated. This issue must be dealt with and corrected as quickly as possible, as the credibility of the data warehouse with the business community is directly related to the quality of the data in the warehouse. This

is true regardless of the fact that many of these quality issues existed previously in the source systems.

Loading substantial amounts of data into the warehouse may require a lot of time. Loads may need to be scheduled for periods in which a sufficient batch window exists if daytime availability is a priority. The amount of data in a data warehouse of moderate size may favor the use of utilities for loading rather than techniques that simply insert data.

While the use of data extraction, transformation, and transfer (ETT) tools is extremely beneficial, the tools are generally complex and require the outlay of additional money to acquire technical training in use of the products during the initial learning curve.

4. Both an integrated political will and budget are required to successfully implement and sustain a data warehouse. The sponsorship and support of the business community are absolutely critical for success. Cultivating and maintaining this sponsorship requires several essential components:

    - a competent business liaison who understands both the data and the business rules and can effectively translate the business needs to the developers;
    - a regular reporting schedule;
    - identification of process and task ownership;
    - clearly defined roles and responsibilities;
    - feasible deadlines, and most importantly;
    - managing users' expectations. Data warehousing requires a new way of thinking, and it is a necessity that the core business users have an understanding of what a data warehouse is, what it is not, and how it may best be employed as a critical business tool.

Appendix D: Farm Service Agency CORE Pilot Data Warehouse

D-1. Overview

In accordance with OMB Circular A-127, the IT Parallel Review Process and the Joint Financial Management Improvement Program (JFMIP) requirements, FSA received technical approval to acquire a commercial off-the-shelf (COTS) accounting software package to satisfy Farm and Foreign Agricultural Service (FFAS) Core Financial Management System requirements. The Core Financial Management System forms the backbone for the FSA's integrated financial management information system. It provides common processing routines and supports common data for critical financial management functions affecting the entire FSA. The changes required to improve Federal financial management systems and support legislated mandates are being implemented by replacing the existing non-compliant FSA/CCC general ledger systems - and business analysis and reengineering of the FSA/CCC business functions.

The Core Accounting System (CORE) initiative will address these requirements and support the vision of an integrated financial management information system. By replacing the CCC legacy financial management systems with the CORE, FSA faces the need to replace hundreds of CCC financial management reports. FSA also has the strategic goal to implement a long-term reporting solution that provides timely, reliable, accurate data, and the ability to analyze financial and program data efficiently, and has the flexibility to respond to changing business requirements, both financial and programmatic.

The FSA pilot data warehouse consists of data from two financial business areas: Debt Management and FSA Salary and Expense and Program General Ledger accounting system (FSA CORE). The debt management data warehouse already existed in a prototype form and was being significantly enhanced at the same time the CCC CORE data warehouse was getting started. The FSA CORE data warehouse is a pilot to the much larger effort of the CCC CORE data warehouse. Therefore, it was decided to merge the debt management data warehouse and the FSA CORE data warehouse into one financial pilot data warehouse. Thus the pilot data warehouse will consist of data from debt management and the FSA salary and expense accounting system. Together, we hope to learn lessons from joining the two data warehouses into a single integrated financial data warehouse.

D-2. CCC CORE Data Warehouse Background

A 5-Year Financial Management Plan was developed by the FSA and the Department in response to the Chief Financial Officer's Act. The goals identified in the Plan include the development of an integrated administrative and financial management information system, enhancing the financial management infrastructure, and improving management accountability. In order to meet the

goals of the 5-Year Plan, the FSA financial and program business areas must be re-engineered into a seamless integrated management information system. As a by-product of the CFO Act, the 5-Year Financial Management Plan provides the charter for the CORE initiative. This plan sends a clear message for USDA's financial managers to change financial management culture, improve management accountability, enhance financial management infrastructure, and improve financial management systems. The plan articulates goals developed by USDA's financial managers:

- Provide high-quality customer service to the Department's internal and external customers;
- Establish framework and environment, and provide timely, accurate financial information to ensure that managers can effectively fulfill their resource management responsibilities and measure performance; assist agencies in implementing strategic planning and performance measures;
- Develop a single, integrated, state-of-the-art administrative and financial management system.

The Congress, Executive Branch, and the Department have sought to improve Federal program and financial systems through a variety of legislative mandates and departmental initiatives. The following legislation and initiatives provide the mandates, guidance, and framework for developing a single, integrated management information system and improving performance measures:

- Chief Financial Officer's (CFO) Act;
- Government Performance and Results Act (GPRA);
- Credit Reform Act;
- Federal Financial Managers Improvement Act (FFMIA);
- Federal Managers' Financial Integrity Act (FMFIA);
- Joint Financial Management Improvement Program (JFMIP);
- OMB Circular A-127, Financial Management Systems;
- USDA Foundation Financial Information System (FFIS);
- USDA Service Center Implementation.

Compliance with the above legislative acts and departmental initiatives will require the Agency to improve, or re-engineer, many business processes to support a performance-driven environment, and to continue to best meet the needs of program customers.

The CORE initiative will address the Federal financial requirements for an agency-consolidated general ledger system, which will replace the existing FFAS general ledger systems. This is the first step towards a single integrated management information system. This initiative was approved by the IRM Review Board and is currently in the implementation phase. The CORE implementation strategy has been to acquire a commercial off-the-shelf (COTS) accounting software package that replaces the existing FFAS general ledger systems. This consolidated general ledger will directly feed the Department's financial system (FFIS) for preparation of the Department's consolidated financial statement and reports, and provide the operational data to prepare the required financial statements for CCC, FSA, Farm Loan Program and Foreign Agricultural Service (FAS).

CORE will provide FSA the opportunity to achieve dramatic improvements in cost, quality, and customer service through the rethinking and redesigning of major business processes. The returns in improved customer service, productivity gains, and overall cost reductions justify the initial expenditure of resources.

FSA's financial management systems track financial events and summarize information to support the mission of FFAS, provide for adequate management reporting, support FSA level policy decisions necessary to carry out fiduciary responsibilities, and support the preparation of auditable financial statements. FSA financial management systems fall into four categories: core financial systems, other financial and mixed systems, shared systems, and departmental executive information systems. These systems must be linked together electronically to be effective and efficient. FSA has developed its own systems architecture consistent with government wide standards and requirements. The following financial management system business areas/types are components of the FSA single, integrated financial management system:

1. Core Financial System – Forms the backbone for the agency's integrated financial management system. It provides common processing routines, supports common data for critical financial management functions affecting the entire FSA, and maintains the required financial data integrity control over financial transactions, resource balances, and other financial systems. The core financial system supports general ledger management, funds management, payment management, receipt management, and cost management. The system received data from other financial systems and from direct user input, and it provides data for financial performance

measurement and analysis and for financial statement preparation.

2. Financial Reporting System - Supports the accumulation and reporting of financial and related information. The system provides information for the annual and other periodic reporting of summary financial and related information including audit trails to systems of original entry and adjustments.

In replacing its financial management systems, FSA faces the need to replace hundreds of existing FSA and CCC reports. FSA also wants to implement a long-term reporting solution that provides the ability to analyze financial data quickly and easily and has the flexibility to respond to changing requirements. FSA had two options to meet these reporting requirements. Option 1 was to produce the reports using the Report Management System (RMS) system that was developed for FSA CORE or option 2, which is to build a data warehouse and use powerful query and reporting tools to generate reports.

Using the RMS system as a benchmark, the estimated costs to develop the reports was between 3 and 4 million dollars and the reports would not provide any analysis capability or the flexibility to respond to change. Therefore, American Management Systems (AMS) and FSA jointly identified option 2 or a data warehouse as a viable solution to meet FSA's needs and at the same time provide a cost effective solution. Thus, the data warehouse has become a critical component of the overall implementation of CORE. In general, a data warehouse extracts information from other systems, integrates and organizes that data for reporting and analysis, and makes the data available directly to managers, analysts, and other end users.

The CORE data warehouse is an architected solution that will manage the central repository of information useful at all levels of an organization (including the executive-, manager-, and analyst-levels) and makes the data accessible through an integrated software and hardware environment. The CORE data warehouse is a hardware and software environment that provides a single point of entry for FSA data, and a consistent, integrated access to enterprise-wide information. The CORE data warehouse will provide the following FSA financial reporting requirements:

1. FSA Financial Management Reporting: The CORE data warehouse will provide financial information in a timely and useful fashion to:

a. support management’s fiduciary role;

b. support the legislative, regulatory and other special management requirements of FSA;

c. support budget formulation and execution functions;

d. support fiscal management of program delivery and program decision making;

e. comply with internal and external reporting requirements, including, as necessary, the requirements for financial statements prepared in accordance with the form and content prescribed by OMB and reporting requirements prescribed by Treasury; and

f. Monitor the financial management system to ensure the integrity of financial data.

2. Performance Measures: The CORE data warehouse will support the Government Performance and Results Act (GPRA) by capturing and producing financial information required to measure program performance, financial performance, and financial management performance required to support budgeting, program management and financial statement presentation. The CORE data warehouse will capture performance measurement data and provide a way of quantifying the relative success or failure of an agency program. A system of performance measures can provide a methodology for focusing on FSA's mission, quantifying the business processes that contribute to the mission, and evaluating among competing priorities for improvement. As new performance measures are established, FSA shall incorporate the necessary information and reporting requirements, as appropriate and feasible, into FSA's financial management systems.

Due to the critical requirement of the data warehouse and the high risk of data warehousing in general, we recommended the development of a pilot data warehouse (Pilot). This strategy agrees with industry standard and with our bench marking of other organizations. The pilot would provide most of the data architecture, establish the infrastructure, institutionalize roles and responsibilities within FSA, provide lessons learned and knowledge transfer that would be used to implement the CCC CORE data warehouse. Because data for the CCC CORE data warehouse will not be available until June 1999, the time frame to build the CCC CORE data warehouse is only four months. It would be project suicide to bypass the pilot data warehouse and attempt to build the data warehouse in a four month period. The data warehouse industry recommends 9 months to build your first data warehouse and to especially avoid a large first-time implementation. The rule is to start small.

The source of data for the pilot project is the FSA CORE accounting data. The data structure and all the reference tables are very close to that of CCC. Therefore, once the pilot is completed most of the software code for the pilot can be reused for the CCC CORE data warehouse. The result is that the four month time frame to build and install the CCC CORE data warehouse can be achieved.

D-3. Background of the Debt Management Data Warehouse

The origin of the debt management data warehouse came as a recommendation from a Business Process Re-engineering project on the Debt Management

reporting process Reporting on debt was inadequate. The quarterly debt portfolio report took 6 weeks and sometimes longer to collect all the data ; data was spread across several systems, and except for the IT people, no one had access to any of the data; analysis of the data was not possible as well as revealing the detail behind the data was not possible. In order to correct this reporting deficiency, a data warehouse was recommended by the BPR team.

Implementation of a debt management data warehouse began with the development of a debt management prototype data warehouse. The development team used existing hardware and software resources. The prototype is in production and proved the value of data warehousing for debt management. However, the prototype did not meet all the requirements of debt management, nor does it take advantage of any query and analysis tools. Because of the reporting requirements of DCIA, management made the decision to launch phase 2 of the debt management data warehouse and expand its capabilities. Phase 2 of the data warehouse has since become part of FSA pilot data warehouse.

D-4. Description of the Pilot Data Warehouse

D-4A. Data Sources:

- FSA salaries and expense accounting data (FSA CORE) - DB2 data base;
- Central Claims Data Base (CCDB) - IDMS data base;
- Claims Referral System (ACAS) - SyBase data base;
- Promissory Notes System - Paradox data base;
- Receivable Data - AE7 Main Frame Generation Data Set - flat files.

D-4B. Hardware

- IBM F50 RSC6000 Data Server;
  - 150 Gigabytes;
  - 4 CPUs;
  - Gigabytes of memory;
- Dell 6300 Web Server.

D-4C. Software

- BrioQuery Enterprise Serve;
- BrioQuery Designer for developers and administrators;
- BrioQuery Explorer for developers;
- BrioQuery Insight for end users;
- SyBase and SyBase IQ for data base.

D-4D. FSA CORE Data Environment

- 15 million row general journal fact table per year;
- FY98 data and FY99 data to date;
- 8 - 12 Gigabytes of data per year;
- 15 dimension tables;
- 5 summary tables.

D-4E. Debt Management Data Environment

- 700 MB of data;
- 10 fact tables - largest fact table has 450,000 rows;
- 9 dimension tables;
- 1 summary table.

D-4F. Data Warehouse Processes

Data is loaded into the data warehouse on a nightly cycle as follows:

1. Data is pushed from the mainframe and other data bases to a staging area on the data server.
2. Data is extracted from the staging area, cleansed, and loaded into the data warehouse.
3. Dimension tables are rebuilt and loaded.
4. Summary tables are built using SQL.
5. The Brio predefined queries are re-generated as the last step.

D-5. Business Problem

The pilot data warehouse is being driven by the following business needs:

1. Detail data about a summary total is not easily accessible.
2. Reconciliation of an account balance is difficult without detail data.
3. Standard reports are not dynamic.
4. Query and analysis is limited.
5. New reports and queries require technical staff or contractor support.
6. Reports can take hours to process.
7. Access to data in the legacy systems is difficult and at times impossible.
8. Financial data originates from several sources making consolidated reporting difficult, time consuming, and labor intensive.
9. Distribution of reports are in hard copy format which means information is not readily available.

D-5. Goals of the Pilot Data Warehouse

The following are the goals of the data warehouse. Each goal however easily translates into a benefit:

- Increase the ease, frequency, and accuracy of reporting;
- Increase the ease, speed, and accuracy of analysis;
- Replace existing hard copy reports with on-line, web-based reporting and query capability;
- Deliver information to anyone with access to the USDA Intranet;
- Lay the foundation and develop plans for full data warehouse development;
- Remove IT support or contractor support from developing new reports;
- Ability to view and explore data not possible in the current legacy environment;
- Establish an infrastructure for future iterations of the CCC Data Warehouse;

- Provide as much re-usability with the tools, data structures, data gathering, and transformation tools and programs used in the course of developing the warehouse;

- Provide a positive Return on Investment (ROI).

Appendix E: FSA Tobacco Data Warehouse Pilot

E-1. Overview

The FSA Tobacco data warehouse pilot was established to provide lessons learned in the methodologies, tools, and organization responsibilities associated with the development and implementation of data warehouses. A special emphasis was directed to the use and evaluation of end user reporting tools.

The pilot is scheduled for completion in late July 1999. The pilot was successful in meeting its objectives. It is the intent of the FSA Tobacco and Peanuts Division to build on this experience by initiating the development of a production Tobacco Data Warehouse. This production data warehouse will support the Tobacco Settlement, Freedom of Information Act requests, program specialists, and other appropriate government and public stakeholders.

The rest of this document describes the Tobacco Data Warehouse products, development techniques, technical architecture, and Lessons Learned.

E-2. Tobacco Data Warehouse Products

Flue-Cured Farm Crop Data for 1997 and 1998 was the primary data loaded into the data warehouse. The data was extracted from SCOAP County File Upload files. This represented about 100,000 records. Additional support files loaded into the data warehouse included Service Center FIPS Code/Names, and Crop Code/Descriptions.

To exercise the BrioQuery reporting tool and to ensure that the data loaded into the data warehouse was accurate - three reports were produced that mirrored existing main frame tobacco reports. These reports are:

- Report of Producer Marketings (PSL-14R);
- Flue-Cured Final Allotment Summary (PSL-15R);
- Size Groups of Tobacco Farms (PSL-17).

Additional ad-hoc reports were produced to learn BrioQuery and its capabilities.

The developed reports as well as the underlying data was made available on the WEB for user access and use. The WEB access function is a capability of the BrioQuery product. A special feature of the WEB page was the inclusion of a data dictionary that described the tables and fields that were available for user use.

E-3. Development Techniques

The following major activities were taken to produce the pilot data warehouse:

- Define user requirements - Joint Application Development sessions
- Identify data sources;
- Determine technical architecture (databases, platforms);
- Extract, cleanse, transform and load data to the data warehouse;
- Train staff on BrioQuery tools;
- Produce reports;
- Place reports on WEB.

E-4. Technical Architecture

The data warehouse was implemented with the following technologies:

- Database: Sybase IQ;
- Reporting Tool: BrioQuery;
- Extraction Tools: COBOL programming language (mainframe);
- Database Platform: IBM F50 RISC/6000;
- BrioQuery Server: Dell 6300;
- Web Server: Dell 6300 (same platform as BrioQuery Server).

E-5. Lessons Learned

Both business and technical staff gained experience and knowledge about a new technology in support of business needs. The lessons learned are grouped under the categories below.

E-5.1 Training

Training was provided in the design of data warehouses, in the use of the BrioQuery reporting tool and in selected database technology. The best experience, however, is the result of actually 'doing it'. This pilot was developed totally in-house without contractor support. One vendor subcontracted training. The facility was not prepared for the course; thus time was lost while the course was rescheduled. Overall training was adequate and effective. After experience is gained on tools, a follow-up session should be arranged with experts. Staff now has a better idea of what questions to ask. After two days of training end-users were able to develop reports of medium complexity. The training would have been even more effective if the test data used in the training course was data associated with the students work area. But with public courses this is not always possible.

E-5.2 Project Management

Standard techniques were used to manage the project. The team kept on schedule in spite of new technology issues and the need to support their current activities. However, when building a production data warehouse, as in any production system, committed resources must not be diverted to other 'priorities'. During key time frames, weekly meeting were held with staff in various locations. Conference calls were effective in keeping staff informed and maintaining action item lists.

E-5.3 Technology

The team was familiar with the hardware platforms and database technology. New combinations of technology created some interesting issues. For example, identifying the correct authorization codes, and associated security levels was time consuming. It requires many different knowledge workers from different functional areas, working together, to figure out access problems. With user groups in two different cities this became even more complex. This pilot provided another opportunity for staff to work together to gain skills and solve problems for the customer. On this project they did it well.

E-5.4 System Performance

With low volumes of data, the system provided adequate response times. On occasion, response was very slow. Better tools are required to quickly pinpoint the bottlenecks. It was not always clear if the database server, BrioQuery server or communication lines were the problem. Although the user can get data in minutes rather than days, expectations rise. Thus, expectation must be managed. As pattern usage is identified, modifications can be made to the data warehouse structure such that response times are minimized for the most frequent queries.

Appendix F: Rural Development Loan Reamortization Pilot Data Warehouse Initiative

F-1 Overview

The Centralized Service Center (CSC) was solicited for a candidate for a pilot data mart that would meet the following criterion: have high demand in the business for a solution; be limited in scope; address a specific business pain; and have an identifiable solution. Using these guidelines, a group comprised of managers and users within CSC selected loan reamortizations as the desired subject area for a pilot data mart.

F-2 Loan Reamortization Data Mart Background

The single most compelling factor in selecting the loan reamortization process was the fact that an increasing number of loans are being reamortized as a means to help customers become successful homeowners; however pertinent information is not readily available for Rural Development program personnel to track the progress of customers whose loans have been reamortized and determine the success and/or failure rates of these reamortizations. Additionally, the work necessary to initiate, process and complete a loan reamortization must be performed by various personnel within multiple branches of the CSC, and the ability to track the work in progress did not exist.

By implementing a pilot data mart for the loan reamortization process, data could be extracted and utilized to satisfy the current and on-going need within CSC to track and analyze loan reamortizations. Applicable information will now be easily available to evaluate key factors related to reamortization processes, including the following:

- whether reamortizations do in fact reduce delinquencies long term, or if reamortizations are a short term fix wherein the customer defaults on the reamortized loan within a relatively short time;
- determine success to failure rate of reamortized customers;
- determine if demographic factors play a role in success or failure rates;
- determine if income to expense ratio currently required for reamortization approval is too high, too low or appropriate;
- determine if particular types of expenses affect success and/or failure rates;

- track the reamortization work processes throughout the various CSC branches to pinpoint any bottlenecks or problem areas;

- provide the ability to evaluate the effectiveness of the origination processor.

The results of these crucial evaluations and determinations will provide CSC program personnel with decision-making information that can be used to restructure, enhance, and optimize the reamortization process.

In addition to the invaluable business benefits described above, the pilot initiative provided the opportunity to glean technical information and garner hands-on experience with data warehousing hardware and software. Benchmark performance measurements and other evaluation parameters yielded technical information that was instrumental in determining the USDA standard suite of Data Warehouse tools.

Implementation of the pilot loan reamortization data warehouse initiative began with the development of an architected prototype data mart that contained standardized data elements, which could serve as common links to any future data marts. The development team used existing hardware and software resources. The prototype is in production and proved the value of data warehousing for analysis of loan reamortizations.

F-3 Description of the Pilot Data Warehouse

F-3A. Data Sources:

- Direct Loan Origination System (DLOS) – 7 flat files created from VSAM data on IBM mainframe.

F-3B. Hardware:

- HP 9000 Series K Data Server:
  - 72 Gigabytes storage;
  - 2 CPUs;
  - 896.8 MB of memory.
- Dell 6300 Web Server.

F-3C. Software

- BrioQuery Enterprise Serve;

- BrioQuery Designer for developers and administrators;
- BrioQuery Explorer for developers;
- BrioQuery Insight for end users;
- Oracle 8.0.5 for data base.

F-3D. Data Warehouse Processes

Data is loaded into the data warehouse on a nightly cycle as follows:

- Data is automatically extracted from the mainframe to a staging area on the data server;
- Data is extracted from the staging area, cleansed, and loaded into the data mart;
- Dimension tables are rebuilt and loaded;
- Summary tables are built;

F-4 Business Problem

The pilot data warehouse is being driven by the following business needs:

- Pertinent information is not readily available for Rural Development program personnel to track the progress of customers whose loans have been reamortized and determine the success and/or failure rates of these reamortizations;
- The work necessary to initiate, process and complete a loan reamortization cannot be tracked throughout multiple branches of the CSC;
- Query and analysis is limited;
- Access to data in the legacy system is difficult and at times impossible.

F-5 Goals of the Pilot Data Warehouse

The following are the goals of the data warehouse. Each goal however easily translates into a benefit:

- Increase the ease, frequency, and accuracy of reporting;

- Increase the ease, speed, and accuracy of analysis;
- Provide ability to view and explore data, which is not currently possible in the existing legacy environment;
- Provide CSC program personnel with decision-making information that can be used to restructure, enhance, and optimize the reamortization process;
- Establish an infrastructure for future iterations of subject area data marts;
- Provide as much re-usability with the tools, data structures, data gathering, and transformation tools and programs used in the course of developing the warehouse;
- Provide a positive Return on Investment (ROI).